Proyecto
Métodos de discretización

1. Introducción

Con este proyecto se buscará comparar el rendimiento de diferentes métodos de discretización mediante el uso del clasificador Naive Bayes en su modo discreto. Para la evalución con el clasificador se implementará la técnica de Cross Validation.

2. Discretizadores

Las técnicas de discretización a utilizar serán:

- Equal Width (NO supervisado): Particionar los datos en `n` particiones iguales de acuerdo al rango de valores de los datos a discretizar.
- Entropy MDL (supervisado): Técnica inventada por Fayyad e Irani es una discretización de arriba hacia abajo, que divide recursivamente el atributo en un corte que maximiza la ganancia de información, hasta que la ganancia es menor que la longitud mínima de descripción del corte.
- CAIM (supervisado): El objetivo del algoritmo CAIM es maximizar la interdependencia clase-atributo y generar un número mínimo de intervalos discretos. El algoritmo no requiere que el usuario predefina el número de intervalos.

Para la aplicación del primer algoritmo de discretización *Equal Width* se utilizará **Weka**, en la que se define un número de 10 particiones (bins). Con el algoritmo *Entropy MDL* se utilizará el set de herramientas de **Orange**, disponibles tanto como aplicación o de manera programática con `Python`. El algoritmo *CAIM* será generado manualmente en `Matlab`, basado en [2].

3. Reporte de resultados

El clasificador de bayes ingenuo se sometió a 10 conjuntos de datos. A todos los datasets se les pasaron las 3 técnicas de discretización.

Para el análisis de los resultados se aplicó la técnica de **Cross Validation**. En esta técnica se particiona el dataset en `k` conjuntos, y en cada `k` iteración se toma una partición como los datos de prueba, y el resto del dataset como entrenamiento.

3.1. Conjunto de datos Iris

Estos datos tienen 5 atributos, donde:

1. Largo del sépalo en cm
2. Ancho del sépalo en cm
3. Largo del petalo en cm
4. Ancho del petalo en cm
5. Clases de iris (Setosa, versicolor y virginica)

Sin discretizar	Equal Width	Entropy MDL	CAIM
27.333 %	71.333 %	93.333 %	94.666 %

Cuadro 1: Precisión del clasificador ingenuo para el conjunto de datos de iris.

Discretizar mejoró de manera significativa el rendimiento del clasificador. CAIM resulto ligeramente mejor de Entropy MDL.

3.2. Conjunto de datos de autenticación de billetes

Estos datos tienen 5 atributos, donde los primeros 4 son datos numéricos que representan las características especiales para un billete, por otro lado, el ultimo atributo es la clase, la cual viene representada por un valor entre 0 y 1, donde 0 indica que no es un billete autentico y 1 el caso contrario.

Se repite la situación de mejora en la clasificación con la discretización, destacando mayormente CAIM y Entropy MDL con rendimientos muy parecidos.

Sin discretizar	Equal Width	Entropy MDL	CAIM
45.401 %	75.985 %	88.613 %	88.901 %

Cuadro 2: Precisión del clasificador ingenuo para el conjunto de datos de autenticación de billetes.

3.3. Conjunto de datos de registros clínicos de insuficiencia cardíaca

Estos datos tienen 13 atributos, donde:

1. edad: edad del paciente (años)
2. anemia: disminución de glóbulos rojos o hemoglobina (booleana)
3. presión arterial alta: si el paciente tiene hipertensión (booleano)
4. creatinina fosfoquinasa (CPK): nivel de la enzima CPK en la sangre (mcg/L)
5. diabetes: si el paciente tiene diabetes (booleano)
6. fracción de eyección: porcentaje de sangre que sale del corazón en cada contracción (porcentaje)
7. plaquetas: plaquetas en la sangre (kiloplaquetas/mL)
8. sexo: mujer u hombre (binario)
9. creatinina sérica: nivel de creatinina sérica en la sangre (mg/dL)
10. sodio sérico: nivel de sodio sérico en la sangre (mEq/L)
11. smoking: si el paciente fuma o no (booleano)
12. tiempo: período de seguimiento (días)
13. Evento de muerte [objetivo]: si el paciente falleció durante el período de seguimiento (booleano)

Si bien se cumplió en mejorar el rendimiento, la clasificación sigue siendo pobre. CAIM resultó con mejor clasificación.

Sin discretizar	Equal Width	Entropy MDL	CAIM
37.241 %	53.793 %	55.172 %	57.314 %

Cuadro 3: Precisión del clasificador ingenuo para el conjunto de datos de registros clínicos de insuficiencia cardíaca.

3.4. Conjunto de datos de diabetes de los indios pimas

Estos datos tienen 9 atributos, donde:

1. Embarazos (Entero)
2. Glucosa (Entero)
3. Presión en sangre (Entero)
4. Grosor de piel (Entero)
5. Insulina (Entero)
6. BMI (Flotante)
7. Función de pedigrí de diabetes (0 al 1)
8. Edad (Entero $>$20)
9. Salida (0,1)

Sin discretizar	Equal Width	Entropy MDL	CAIM
41.315 %	67.631 %	70.526 %	68.447 %

Cuadro 4: Precisión del clasificador ingenuo para el conjunto de datos de diabetes de los indios pimas.

3.5. Conjunto de datos de riesgo de comportamiento del cáncer de cuello uterino

Estos datos tienen 20 atributos (Enteros), donde:

1. comportamiento comer
2. comportamiento higienepersonal
3. intención agregación
4. intención compromiso
5. actitud consistencia
6. actitud espontaneidad
7. persona significativa norma
8. norma cumplimiento
9. percepción vulnerabilidad
10. percepción severidad
11. motivación fuerza
12. motivación voluntad
13. apoyo social emocionalidad
14. apoyo social apreciación
15. apoyo social instrumental
16. empoderamiento conocimiento
17. empoderamiento habilidades
18. empoderamiento deseos
19. ca cervix (este es un atributo de clase, 1 = tiene cáncer de cuello uterino, 0 = no tiene cáncer de cuello uterino)

3.6. Conjunto de datos sonar

Los datos de este conjunto contiene 111 patrones obtenidos al hacer rebotar señales de sonar en un cilindro de metal en varios ángulos y bajo diversas condiciones. El ultimo atributo pertenece a la clase, de la cual solo hay 2 valores, Mina o Roca.

Sin discretizar	Equal Width	Entropy MDL	CAIM
55.714 %	57.142 %	58.561 %	77.142 %

Cuadro 5: Precisión del clasificador ingenuo para el Conjunto de datos de riesgo de comportamiento del cáncer de cuello uterino.

Sin discretizar	Equal Width	Entropy MDL	CAIM
47.5 %	52 %	56.5 %	60 %

Cuadro 6: Precisión del clasificador ingenuo para el Conjunto de datos de .

3.7. Conjunto de datos de ionosfera

Estos datos tienen 34 atributos continuos y el atributo clase(35), es "bueno.º "malo".

Sin discretizar	Equal Width	Entropy MDL	CAIM
63.428 %	81.428 %	87.142 %	86.517 %

Cuadro 7: Precisión del clasificador ingenuo para el Conjunto de datos de ionosfera.

3.8. Conjunto de datos de identificación de vino

Estos datos tienen 13 atributos continuos y el atributo clase, donde:

1. alcohol
2. acido málico
3. restos
4. alcalinidad de los restos
5. magnesio
6. fenoles totales

7. flavonoide
8. no-Flavonoides fenoles
9. proantocianidina
10. intensidad de color
11. matiz
12. OD280/OD315
13. prolinas
14. Variable de salida: Identificación de tipo de vino (1-3).

Sin discretizar	Equal Width	Entropy MDL	CAIM
7.647 %	55.882 %	78.235 %	86.471 %

Cuadro 8: Precisión del clasificador ingenuo para el Conjunto de datos de vino.

3.9. Fallos de simulación de modelos climáticos

Este conjunto de datos contiene registros de fallas de simulación encontradas durante conjuntos de cuantificación de incertidumbre (UQ) del modelo climático. El objetivo es predecir los resultados de la simulación del modelo climático (fallido o exitoso) dados los valores escalados de los parámetros de entrada del modelo climático.

Sin discretizar	Equal Width	Entropy MDL	CAIM
9.444 %	48.148 %	60.555 %	64.921 %

Cuadro 9: Precisión del clasificador ingenuo para el Conjunto de datos de fallas de simulación de modelos climáticos.

3.10. Conjunto de datos de creditos de banco alemán

Estos datos tienen 20 atributos continuos y el atributo clase. Este dataset recolecta datos sobre sus estados de cuenta, estados laborales, estatus marital, tipo de vivienda, características de la familia e información relevante sobre el prestamo solicitado por el cliente. Se clasifican a los que se les concedería un crédito y a los que no.

Sin discretizar	Equal Width	Entropy MDL	CAIM
53.8 %	55.2 %	58.6 %	62.2 %

Cuadro 10: Precisión del clasificador ingenuo para el Conjunto de datos de créditos de banco alemán.

3.11. Conjunto de datos de clasificación de semillas

De acuerdo a ciertos parámetros, se busca determinar el tipo de semilla al cuál pertenece. Los atributos son:

1. area de la semilla
2. perímetro de la semilla
3. compacidad
4. largo del nucleo
5. ancho del nucleo
6. coeficiente de asimetría
7. longitud del surco del grano
8. Variable de salida: Tipo de grano (1-3).

Referencias

[1] David Aha y otros. *UCI Machine Learning Repository*. 1987. URL: `https://archive.ics.uci.edu/ml/datasets.php`.

Sin discretizar	Equal Width	Entropy MDL	CAIM
11.579 %	79.473 %	85.78 %	90 %

Cuadro 11: Precisión del clasificador ingenuo para el Conjunto de datos de semillas.

[2] Lukasz A. Kurgan y Krzysztof J. Cios. "CAIM Discretization Algorithm". En: *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* 16.2 (feb. de 2004), págs. 145-153.

[3] Jason Brownlee. *Naive Bayes Classifier From Scratch in Python*. 2019. URL: `https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/`.

[4] gsBatra. *NaiveBayesClassifier*. 2021. URL: `https://github.com/gsBatra/NaiveBayesClassifier/blob/master/NaiveBayesClassifier.py`.

[5] SidhanthaPoddar2. *Binning in Data Mining*. 2022. URL: `https://www.geeksforgeeks.org/binning-in-data-mining/`.

[6] Orange Visual Programming. *Orange - Discretize*. URL: `https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/data/discretize.html`.